OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE

Cosmetic

Extending the Range of Measurement for Minimally Invasive Treatments by Adding New Concepts to FACE-Q Aesthetics Scales

Anne F. Klassen, DPhil[*]
Andrea L. Pusic, MD[†]
Manraj Kaur, PhD[‡]
Charlene Rae, PhD[*]
Lotte Poulsen, MD, PhD[§]
Jasmine Mansouri, BSc[¶]
Elena Tsangaris, PhD[‡]
Steven Dayan, MD[||]
Jennifer Klok, MD[**]
Kathleen Armstrong, MD[††]
Katherine Santosa, MD[‡‡]
Stefan Cano, PhD, CPsychol, AFBPsS[§§]

Background: The Satisfaction with Face Overall and Psychological Function scales are the most frequently used FACE-Q Aesthetics module scales. This study aimed to extend their range of measurement by adding and testing new concepts. We aimed to create FACE-Q Aesthetics item libraries.

Methods: In-depth concept elicitation interviews were conducted. Concepts were formed into items and refined through multiple rounds of patient and expert input. The items were tested with people living in the United States, Canada, and the United Kingdom who had minimally invasive facial aesthetic treatments. Participants were recruited through an online platform (ie, Prolific). Psychometric properties were examined using Rasch measurement theory analysis, test–retest reliability, and construct validity.

Results: We conducted 26 interviews. New concepts were developed into items and refined with input from 12 experts, 11 clinic patients, and 184 Prolific participants. A sample of 1369 Prolific participants completed 52 appearance and 22 psychological items. After removing 10 and 2 items respectively, the psychometric tests provided evidence of reliability with the person separation index, Cronbach alpha, and test–retest reliability values without extremes of 0.88 or more. For validity, lower scores were associated with looking older than one's age, being more bothered by facial skin laxity, treatment wearing off, and having deeper lines on Merz Assessment scales. Short-form scales formed from the 42 appearance items provide examples of item library application.

Conclusions: This study provides an innovative means to customize scales to measure appearance and psychological function that maximizes content validity and minimizes respondent burden in the context of minimally invasive treatments. *(Plast Reconstr Surg Glob Open 2024; 12:e5736; doi: 10.1097/GOX.0000000000005736; Published online 10 April 2024.)*

*From *McMaster University, Hamilton, Ontario, Canada; †Patient Reported Outcome, Value, and Experience (PROVE) Center, Department of Surgery, Division of Plastic and Reconstructive Surgery, Brigham and Women's Hospital, Boston, Mass.; ‡Patient Reported Outcome, Value, and Experience (PROVE) Center, Brigham and Women's Hospital, Boston, Mass.; §Research Unit for Plastic Surgery, University of Southern Denmark, Odense, Denmark; ¶Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada; ||Denova Research, Chicago, Ill.; **Dr. Jennifer Klok clinic, West Vancouver, British Columbia, Canada; ††McLean Clinic, Mississauga, Ontario, Canada; ‡‡Artisan Plastic Surgery, Atlanta, Ga.; and §§Modus Outcomes (a company of Thread), Cheltenham, United Kingdom.*

Received for publication January 10, 2024; accepted February 28, 2024.

DOI: 10.1097/GOX.0000000000005736

INTRODUCTION

The US Food and Drug Administration defines patient-reported outcome measures as questionnaires that measure how patients function and feel by asking them directly.[1] Such tools are increasingly used to promote shared decision-making in patient care, as quality metrics, and in comparative effectiveness research.[2–4]

Our research team developed a modular PROM to evaluate outcomes for surgical and nonsurgical facial aesthetic treatments, ie, FACE-Q Aesthetics.[5–12] This PROM is composed of 40 separate scales and checklists that measure satisfaction with appearance, adverse effects, and health-related quality of life. FACE-Q Aesthetics has been used extensively to evaluate the safety and effectiveness of facial aesthetics treatments.[13,14] A recent systematic

Disclosure statements are at the end of this article, following the correspondence information.

Related Digital Media are available in the full-text version of the article on www.PRSGlobalOpen.com.

review showed that 114 studies had used one or more FACE-Q Aesthetics scales, and that the Satisfaction with Face Overall[6] and Psychological Functional[9] scales were the most frequently used (ie, 52 studies and 45 studies, respectively).[15] In 2020, both scales were qualified as a medical device development tool by the USA Food and Drug Administration, with Face Overall recommended as a co-primary or secondary end point in clinical trials, and Psychological Function as an ancillary end point.[16]

Since FACE-Q Aesthetics was developed, the global medical aesthetics market has expanded dramatically.[17] As more people access an increasing range of treatments to rejuvenate appearance and improve health-related quality of life, it is vital that outcomes are carefully evaluated. Standard practice for PROM design involves the development of static forms that comprise a relatively short number of questions for use in a specific context of use. More recently, PRO item libraries[18,19] and item banks[20–22] allow researchers to pick a subset of items to maximize content validity and ensure the PROM is fit-for-purpose. Currently, no item libraries or banks are available for use in facial aesthetics.

The specific aims of our study were (1) to elicit concepts important to measuring facial appearance and psychological function for minimally invasive treatments; (2) to develop, refine, and test items to extend the range of measurement for the Face Overall and Psychological Function scales, and (3) to provide example short-form scales to show how the items can be used.

METHODS

Research Ethics

This study was coordinated at McMaster University (Canada). Ethics board approval (approval no.: 13603) was obtained from the Hamilton Integrated Ethics Board (Canada) before commencing the study.

Approach

We used a mixed methods approach[23] and followed international guidelines for PROM development.[1,24–26] Figure 1 shows the methods we followed. The qualitative phase used interpretative description.[27] Between 22 October 2021 and 31 March 2022, participants were recruited from three plastic surgery clinics in Canada and three in the United States. Clinic staff were asked to recruit people who varied by age, gender, race, and treatment type. Patients who agreed to an interview were contacted by a researcher who explained the study and obtained informed consent.

Interviews were conducted by an experienced qualitative interviewer by phone or using a secure web conferencing platform (ie, Zoom). Supplemental Digital Content 1 shows the topics covered. (**See table, Supplemental Digital Content 1,** which displays qualitative interview topic guide. **http://links.lww.com/PRSGO/D145.**)

All interviews were audio-recorded, transcribed, and coded by labeling concepts with a domain and major/minor theme. Transcripts were coded independently by two coders, who achieved consensus on discrepancies. Codes were transferred to MS Excel and refined through constant comparison.[28] Interviews continued until saturation of most concepts was reached.[29] A thank-you gift card of $100 was provided to participants.

Takeaways

Question: Which concepts are important to patients for measuring facial appearance and psychological function in the context of minimally invasive facial treatments?

Findings: In our large international mixed methods study, we were able to extend the range of measurement for two key FACE-Q Aesthetic scales to include 42 appearance items and 20 psychological items and to provide example short-form scales.

Meaning: We provide an innovative means to customize scales to measure facial appearance and psychological function in facial aesthetics research and clinical care.

Scale Development and Refinement

An item pool was developed and refined through several steps.

First, in October 2022, clinic participants were invited to provide feedback in REDCap.[30] To examine comprehension and relevance, participants selected one answer from four options: (1) I do not understand the question; (2) I understand the question, but it could be worded better; (3) I understand the question, but it is not relevant to me; and (4) I understand the question and it is relevant to me. For comprehensiveness, an open text box was provided. Items identified as problematic were dropped or revised. Participants received a $30 thank-you gift card.

Second, cognitive debriefing interviews were performed on Zoom with an experienced interviewer. Participants provided feedback on instructions, items, and response options, and suggested missing content. Interviews were audio-recorded, transcribed, and analyzed. Participants received a $70 thank-you gift card. In addition, clinical experts and representatives from the aesthetics industry were invited by email to highlight items they deemed not relevant to patients and to suggest missing concepts.

Third, content validity was explored in a larger sample using an online crowd working platform [ie, Prolific (www.prolific.co)]. A screening survey was conducted in December 2022 for Canada and the United States, and August 2023 for the United Kingdom. The number of residents fluent in English for the Canada/US sample was 121,170 and for the UK sample was 37,458. Participants were paid the equivalent of 10.80 GBP per hour. We included from the sample people who in the past 12 months had one or more of the treatments described in Supplemental Digital Content 2 and excluded anyone who had not been to a plastic surgery or dermatology clinic for treatment in the past 12 months, and anyone who chose "none" or "other" for treatment type. (**See table, Supplemental Digital Content 2,** which displays screening questions used in Prolific. **http://links.lww.com/PRSGO/D146.**)

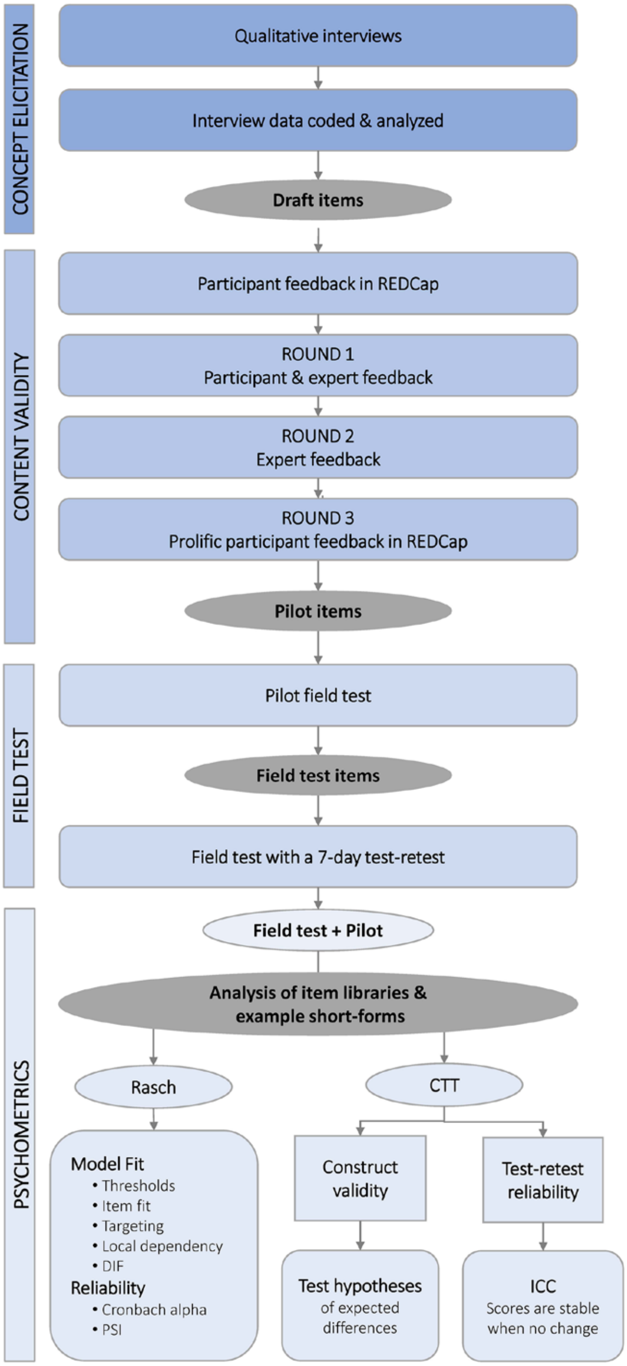

Fig. 1. Methods flow diagram.

Participants were invited to read each item and choose one answer from the following: (1) I do not understand the question; (2) I understand the question, but it is not relevant to me; and (3) I understand the question and it is relevant to me. Open text boxes were provided for missing concepts.

In February 2023, we conducted a pilot field-test using the initial US/Canada Prolific sample. The pilot field-test data were examined to identify and remove items with extreme misfit to the Rasch model.

Following the pilot, for the field-test, in March 2023, we recruited a sample of Prolific people from Canada and the United States. To closely match ASPS age statistics for people having aesthetic treatments,[17] in August 2023 we recruited an older Prolific field-test sample from Canada, the United States and the United Kingdom. Supplemental Digital Content 2 inclusion criteria were used.

Data were downloaded into SPSS, version 28 (IBM Corporation, Armonk, New York, N.Y.) and imported into RUMM2030 software[31] for Rasch measurement theory (RMT) analysis.[32,33] For the RMT analyses, we started with the 10 items that formed the original scales and included as many additional items as fit the Rasch model. The unrestricted Rasch model for polytomous ordered responses was used. Table 1 shows the psychometric tests performed.

RESULTS

Concept Elicitation and Scale Refinement

Characteristics of the study samples are shown in Tables 2 and 3. Coding and analysis of the 26 qualitative interviews identified 57 concepts relevant to measuring facial appearance and 25 relevant to psychological function. Eleven of the 26 interview participants provided feedback in REDCap. Of the 627 ratings (ie, 11 participants × 57 items), 0.5% of ratings were “I do not understand”; 3% of ratings were “I understand this question, but it could be worded better; 9.7% of ratings were “I understand this question, but it is not relevant to me”; and 86.8% of ratings were “I understand this question and it is relevant to me.” For the 25 psychological items, of 275 ratings (ie, 11 participants × 25 items), 0 ratings were “I do not understand”; 0.4% of ratings were “I understand this question, but it could be worded better; 9.8% of ratings were “I understand this question, but it is not relevant to me”; and 89.8% of ratings were “I understand this question and it is relevant to me.”

Supplemental Digital Contents 3 and 4 show the item-level decisions (ie, retain, revise, drop, add) made in each round of refinement. (**See table, Supplemental Digital Content 3,** which displays changes made to face items in each round sorted by Prolific sample relevance ratings. **http://links.lww.com/PRSGO/D147.**) (**See table, Supplemental Digital Content 4,** which displays changes made to psychological items in each round sorted by Prolific sample relevance ratings. **http://links.lww.com/PRSGO/D148.**)

In round 1, seven cognitive debriefing interviews were performed, and feedback was obtained from three aesthetic plastic surgeons and one plastic surgery resident from Canada. No changes were made to the psychological items. For the appearance items, 51 items were retained, four items were revised, and two items were dropped, resulting in 55 items.

Round 2 included five plastic surgeons, one dermatologist and two industry experts from Denmark, Canada, Sweden, and the United States. In this round, no changes were made to the psychological items. For

Table 1. Psychometric Tests Performed and Their Interpretation

Test	Description
Thresholds for item responses	Item response options need to be ordered on a continuum (eg, a score of 1 lower than a score of 2). This approach is used to create a hierarchy of items to determine how items were ordered from easiest to hardest to endorse.
Item fit	The extent to which observed data fit expected values based on the Rasch model. Item fit was assessed by inspecting fit residuals and chi-square statistics. Fit residuals summarize the observed and expected responses to an item and should ideally lie within ±2.5. Items should have chi-square values that are nonsignificant after Bonferroni adjustment. For the item fit analysis, the sample size was amended to 500 to adjust the *P* values given the large sample.[33]
Local dependency	Residual correlations were examined to identify any greater than 0.30 above the average correlations. Subtest analysis was performed to determine the impact of local dependency on scale reliability.[34]
Scale-to-sample targeting	Targeting looks at the spread of person locations (eg, satisfaction with face) against the spread of item locations (eg, range of measurement). A scale that is better targeted has more coverage with the mean person location close to the center of the scale.[35] We also computed the proportion of the sample that scored on scale.
Differential Item Functioning (DIF)	DIF examined the extent to which items were invariant across age (ie, 20–29, 30–39, 40–49, ≥50), gender (women versus men), and country (United States, Canada, United Kingdom). The sample was amended to 500 to adjust the *P* values for the large sample. Random samples with equal size samples in subgroups were chosen. When potential DIF was identified, variables were split for the relevant items, with both original and split person locations correlated to examine the impact of DIF on scale scoring.[36] The analysis was repeated three times to determine if the results were stable.
Reliability	1. Person separation index: this statistic determined how well people in the sample were separated by the scale items.[37] 2. Cronbach alpha: this statistic was used to examine internal reliability. 3. Test–retest reliability: a subset of participants completed the survey twice. We excluded anyone who reported an important change in satisfaction with face and in psychological function, or who completed the TRT outside of 7–14 days. We examined extremes using boxplots. Intraclass correlation coefficients were computed with a two-way random effects model with and without extremes included. Reliability values should be >0.70.[38,39]
Construct validity	Rasch logit scores were transformed into 0 (worse) to 100 (best), and short forms scores were calibrated using the item-bank approach. Parametric or nonparametric tests were used depending on the distribution of the data. Statistical significance was set at a two-tailed *P* value of <0.05. 1. Scores would be incrementally lower for participants who reported they looked older on the FACE-Q Aesthetics Age Visual Analogue Scale.[7] This scale measures how many years younger or older people think they look compared with their actual age (range from ±15 years). Scores were categorized as follows: look younger, look age, and look older. 2. Scores would be incrementally lower based on how much participants report their aesthetic facial treatment(s) has worn off (not at all, partially, completely). 3. Scores would be incrementally lower based on how much (not at all, a little, moderately, very, extremely) participants were bothered by lax or loose skin on their face. 4. Scores would be incrementally lower based on the depth (none, mild, moderate, severe/very severe) of dynamic (ie, crow feet, forehead lines, glabellar lines) and static (ie, nasolabial folds, marionette lines, lip lines) self-reported Merz Assessment Scale scores.[40] The Merz Assessment Scale is a validated and reliable photonumeric scale used to rate severity of facial lines using five categories.

TRT, test–retest reliability.

the appearance items, 55 items were retained, and two items were added.

In round 3, a total of 556 US and Canadian Prolific participants accessed the cognitive screening survey. The 194 people who met the study criteria were invited to complete the survey; 156 did, and of these, 144 met the inclusion criteria. A total of 122 participants completed the 57 face items, and 144 completed the 25 psychological items. A total of 40 participants from the UK sample were invited to complete the cognitive survey before the field test, increasing the sample to 184. For the face items, of 9234 ratings (ie, 162 participants × 57 items), the option “I do not understand the question” was chosen 1.6% of the time and the option “I understand the question and it is relevant to me” was selected 73.5% of the time. For the psychological items, of the 4600 ratings (ie, 184 participants × 25 items), “I do not understand the question” was chosen 1.1% of the time and the option “I understand the question and it is relevant to me” was selected 77.3% of the time. From Face, we dropped four items and revised one, and from Psychological, we dropped three items.

From the Canada/US cognitive sample, all 144 participants were invited to complete a pilot field-test, and 110 did. Based on the RMT analysis, one appearance item was dropped. The field-test version included 52 face items and 22 psychological items.

Psychometric Analyses

A total of 4301 Prolific participants responded to the screening surveys. We removed 1365 duplicates, incompletes, and ineligibles. Of the remaining participants, 1895 met the inclusion criteria and were invited to complete the survey. Of these, 1458 responded. From these, we excluded 199 respondents as follows: incomplete (N = 95), no treatment (N = 84), reported “other” for the type of treatment (N = 14), and unreliable answers (N = 6).

The field-test sample had 1259 eligible participants. We included data for the 110 pilot field-test participants, providing 1369 participants for the RMT analysis. Table 4 shows scale-level results, and Supplemental Digital Contents 5 and 6 show the item RMT and DIF results. (**See table, Supplemental Digital Content 5,** which displays RMT item-level fit statistics and DIF results for Face item bank/library. **http://links.lww.com/PRSGO/D149**.) (**See table, Supplemental Digital Content 6**, which displays RMT item-level fit statistics and DIF results for the psychological item set. **http://links.lww.com/PRSGO/D150**.)

Facial Appearance

In the first step, data for the original 10-item Face Overall scale[6] fit the Rasch model [$\chi^2 = 74.4$, degrees of freedom (df) = 90, $P = 0.88$]. All 10 items had ordered

Table 2. Participant Characteristics

			Prolific			
		Qualitative Sample	Cognitive Sample		Psychometric Sample	
		N = 26	N = 184	%	N = 1369	%
Country	Canada	6	21	11.4	107	7.8
	United Kingdom	0	40	21.7	721	52.7
	United States	20	123	66.8	540	39.4
	Missing	0	0.0	0.0	1.0	0.2
Age	20–29	3	39	21.2	229	16.7
	30–39	6	43	23.4	308	22.5
	40–49	7	45	24.5	445	32.5
	50–59	6	37	20.1	255	18.6
	≥60	4	20	10.9	132	9.6
Sex	Women	23	154	83.7	1005	73.4
	Men	3	28	15.2	351	25.6
	Gender diverse	0	2	1.1	10	0.8
	Prefer to not answer	0	0	0	3	0.2
Race	White	22	139	75.5	1048	76.6
	Black	2	9	4.9	96	7.0
	Latin American	0	7	3.8	35	2.6
	East Asian	0	6	3.3	45	3.3
	Middle Eastern	0	3	1.6	10	0.7
	South Asian	1	5	2.7	40	2.9
	Southeast Asian	1	2	1.1	12	0.9
	Indigenous	0	1	0.5	1	0.1
	Mixed race	0	10	5.4	68	5.0
	Other/missing/prefer to not answer	0	2	1.0	14	1.1
Marital Status	Married/common law	16	88	47.9	753	55.0
	Single	7	60	32.6	428	31.3
	Divorced	2	23	12.5	122	8.9
	Separated	0	7	3.8	35	2.6
	Widowed	1	3	1.6	14	1.0
	Other/missing/prefer to not answer	0	3	1.6	17	1.2
Fitzpatrick Skin Type	Always burn and never tan	2	10	5.4	98	7.2
	Usually burn and minimally tan	9	45	24.5	371	27.1
	Mild burn and then tan	9	81	44.0	503	36.7
	Rarely burn and always tan	4	25	13.6	264	19.3
	Rarely burn and tan very easily	1	16	8.7	111	8.1
	Never burn and never tan	1	4	2.2	22	1.6
	Missing		3	1.6	0	0
Highest Education	Some high school	0	3	1.6	7	0.5
	High school	1	10	5.4	113	8.3
	Some college, trade, or university	4	25	13.6	200	14.6
	College, trade, or university degree	9	103	56.0	690	50.4
	Some masters or doctoral degree	0	7	3.8	86	6.3
	Masters or doctoral degree	11	36	19.6	272	19.9
	Missing/prefer to not answer	1	0	0	1	0.1

thresholds and good item fit to the Rasch model with nonsignificant chi-square *P* values after Bonferroni adjustment. Fit residuals were within ±2.5 for seven items. Reliability was high: person separation index (PSI) and Cronbach alpha values were 0.90 or more. Figure 2 shows the person-item threshold distribution. The bottom histogram mapped out the range of measurement for satisfaction with appearance, and the top histogram shows the sample (ie, 95% scored on the scale).

In the next step, the best solution that included the 10 original items incorporated 42 additional items. Data fit the Rasch model (($\chi^2 = 406.2$, df = 378, $P = 0.15$). All 42 items had ordered thresholds (Fig. 3), and 41 items had nonsignificant chi-square *P* values after Bonferroni adjustment. Fit residuals were within ±2.5 for 21 items. DIF was evident for seven items for age-group and three items for gender. Correlations between the person locations before and after splitting for DIF for the relevant patient characteristics showed no impact on scoring (r = 1.00). Reliability was high with PSI and Cronbach alpha values greater than or equal to 0.97. Eighteen pairs of items had residual correlation values of more than 0.30, suggestive of local dependency. After subtests were performed, there was little impact on reliability: PSI

Table 3. Treatment History Reported by the Qualitative Sample and Prolific Participants

			Prolific			
		Qualitative Sample	Cognitive Sample		Psychometric Sample	
		N = 26	N = 184	%	N = 1369	%
Injectable	Botox	18	124	67.4	592	43.2
	Filler	17	121	65.8	394	28.8
	Platelet-rich plasma	1	13	7.1	58	4.2
	Skin booster	0	17	9.3	115	8.4
Skin Resurfacing	Microdermabrasion	7	81	44.0	496	36.2
	Chemical peel	16	74	40.2	488	35.6
	Hydrafacial	2	65	35.3	578	42.2
	Laser	14	47	25.5	234	17.1
	Microneedling	2	46	25.0	297	21.7
	Light therapy	14	40	21.7	205	15.0
Skin Tightening	Radiofrequency	7	21	11.4	151	11.0
	High intensity ultrasound	0	17	9.2	138	10.1
	Thread lift	1	13	7.1	96	7.0
Fat Removal	Fat removal	1	14	7.6	86	6.3

Table 4. RMT Scale-level Statistics and Other Psychometric Results

												Test–Retest Reliability			
								PSI		α				95% CI	
Scale	Items N	Sample N	RMT N	Score on Scale %	χ^2	DF	*P* Value	+Ext	–Ext	+Ext	–Ext	N	ICC – /+ Extremes	Lower Bound	Upper Bound
Face Overall[6]	10	1369	1300	95.0	74.39	90	0.88	0.92	0.90	0.93	0.91	97	0.90	0.84	0.93
												107	0.76	0.66	0.84
Face Item Library	42	1369	1334	97.4	406.23	378	0.15	0.97	0.98	0.98	0.98	97	0.88	0.83	0.92
												107	0.74	0.62	0.82
Facial Rejuvenation	10	1369	1287	94.0	112.31	90	0.06	0.94	0.93	0.96	0.94	97	0.88	0.82	0.92
												107	0.76	0.65	0.84
Facial Appearance	10	1369	1307	95.5	74.91	90	0.87	0.92	0.91	0.93	0.92	97	0.89	0.84	0.93
												107	0.76	0.64	0.83
Facial Aging	10	1369	1299	94.9	60.28	90	0.99	0.91	0.90	0.93	0.92	97	0.89	0.84	0.93
												107	0.77	0.66	0.84
Psychological Function[9]	10	1369	1156	84.4	170.72	90	0.00	0.95	0.94	0.97	0.95	102	0.90	0.84	0.93
												109	0.83	0.75	0.88
Psychological Item Library	20	1368	1232	90.1	231.20	180	0.01	0.97	0.97	0.98	0.98	101	0.87	0.85	0.93
												109	0.83	0.74	0.88

α, Cronbach alpha; +ext, with extremes; -ext, without extremes; ICC, intraclass correlation coefficient; CI, confidence intervals.

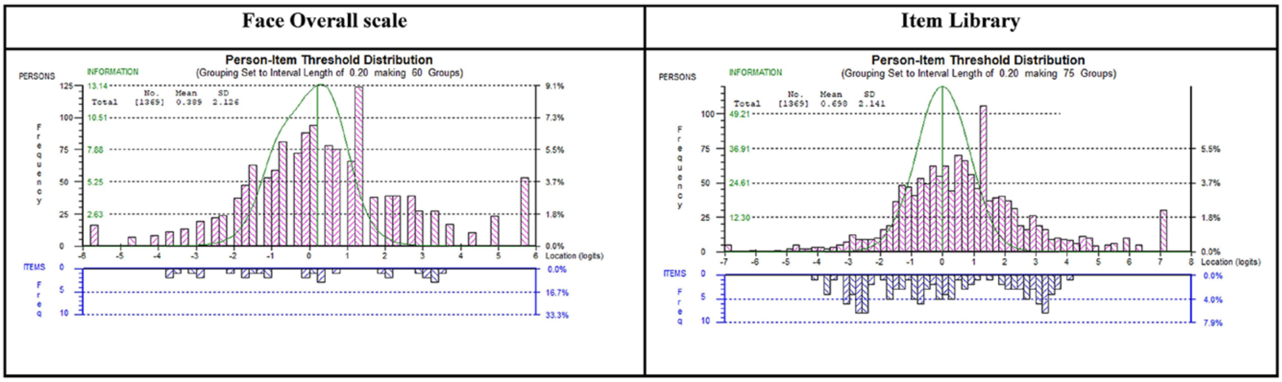

Fig. 2. Person-item threshold distribution for the FACE-Q Aesthetics Face Overall scale[6] and face item library.

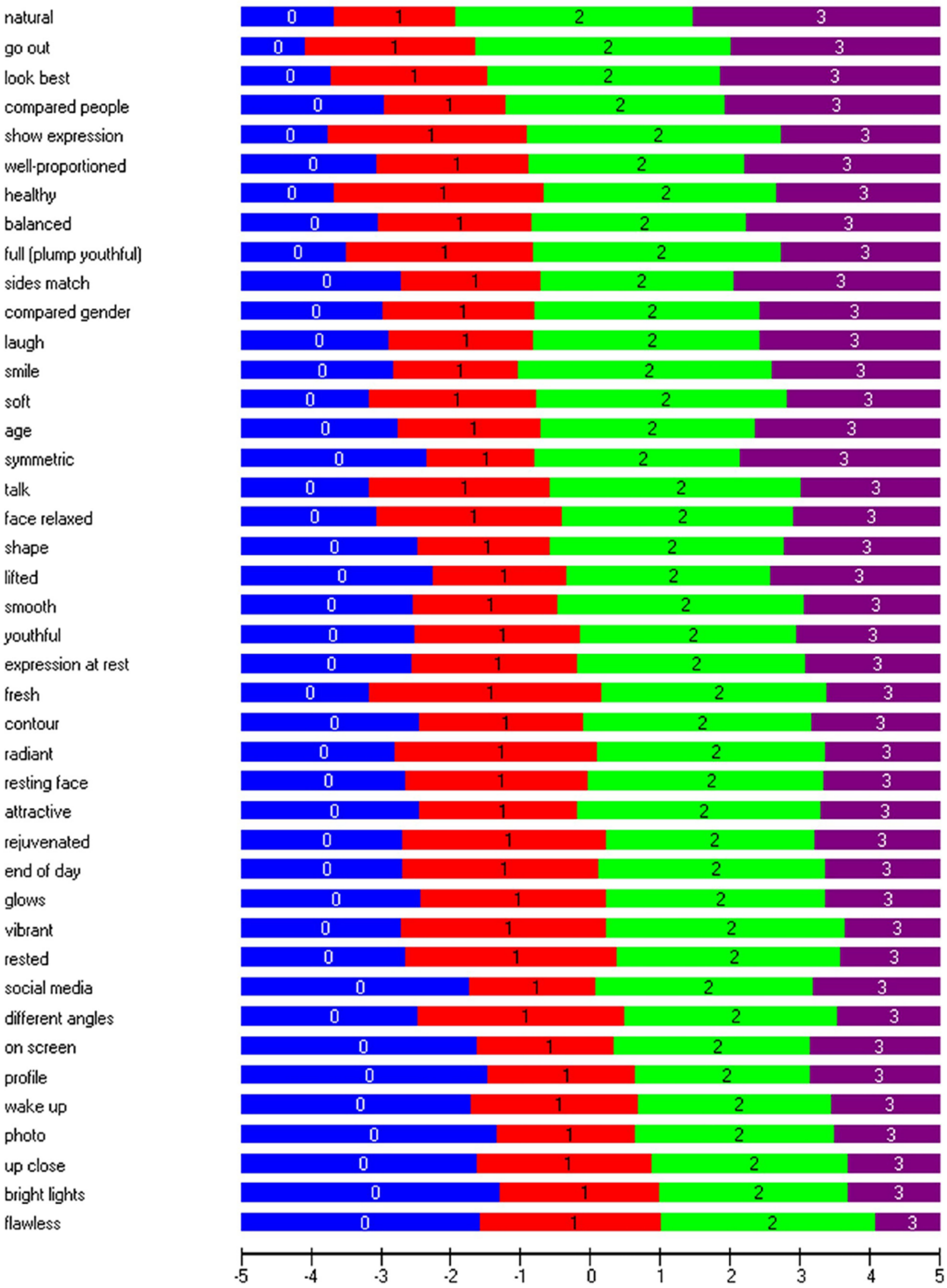

Fig. 3. Threshold map for the FACE-Q Aesthetics face item library.

values = 0.97 and Cronbach alpha values = 0.96. Figure 2 shows the person-item threshold distribution; 97.4% of participants scored within the range of measurement provided by the scale. Floor (0.4%) and ceiling (2.2%) effects were low.

Three example 10-item short forms were tested: Facial Rejuvenation (eg, how fresh, youthful, and radiant the face looks), Facial Appearance (eg, on a screen, photographs, bright lights), and Facial Aging (eg, how lifted, full and age the face looks). Data fit the Rasch models for each scale (Table 4). All 10 items in each scale had ordered thresholds and good item fit to the Rasch model with nonsignificant *P* values after Bonferroni adjustment. Reliability was high with PSI and Cronbach alpha values of 0.90 or more. One pair of items in the Facial Rejuvenation scale evidenced local dependency; after a subtest the PSI and Cronbach alpha values were 0.93 or more. The scales were well targeted to the sample: 94% or more scored on the scale.

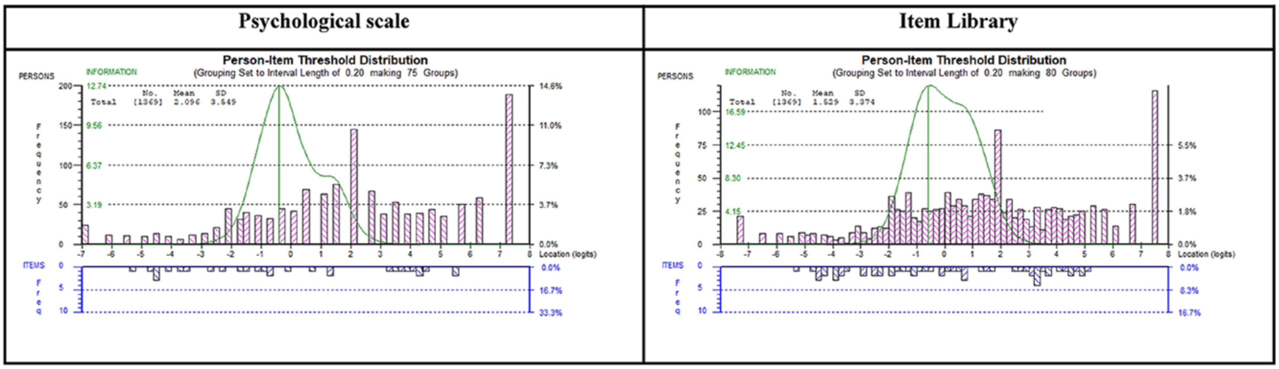

Fig. 4. Person-item threshold distribution for the FACE-Q Aesthetics Psychological Function scale[9] and psychological item library.

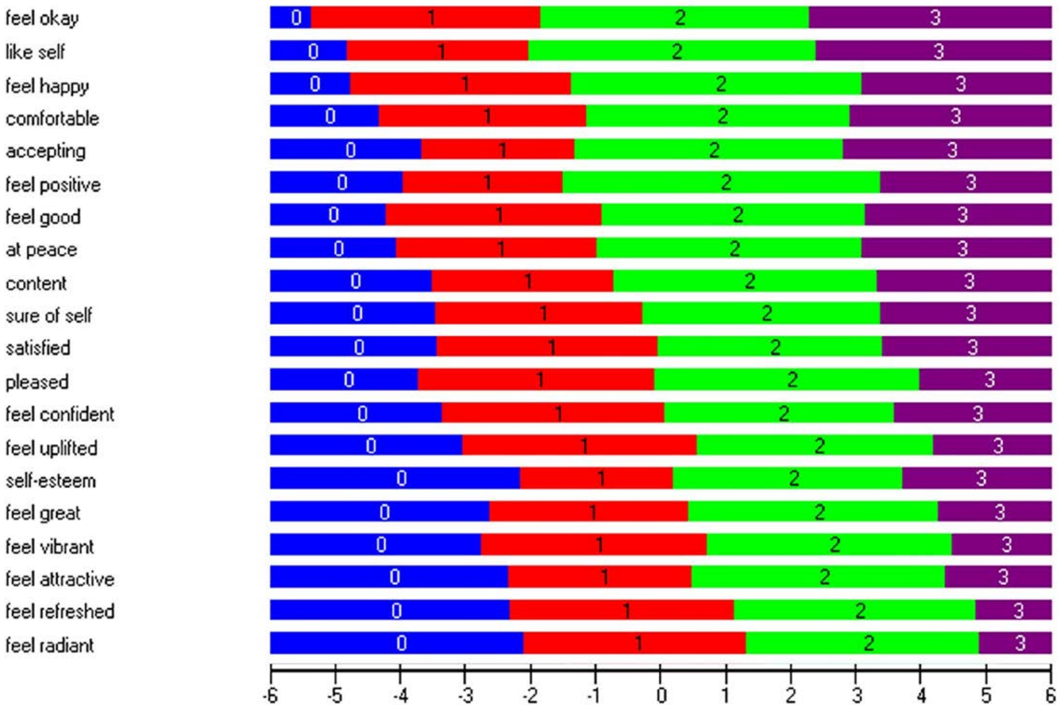

Fig. 5. Threshold map for the FACE-Q Aesthetics psychological item library.

Psychological

In the first step, data for the original 10-item Psychological scale[9] evidenced some misfit to the Rasch model ($\chi^2 = 84.4$, df = 170.7, $P < 0.001$) due to one misfit item ("I feel attractive"). However, all 10 items had ordered thresholds, nine items fit the Rasch model with nonsignificant chi-square P values after Bonferroni adjustment, and item fit was with ±2.5 for six items. Reliability was high with PSI values and Cronbach alpha values greater than or equal to 0.94. Targeting was good; 84.4% of the sample scored on the scale (Fig. 4).

In the next step, the best solution that included the 10 original items incorporated 10 additional items. Data evidenced marginal misfit to the Rasch model ($\chi^2 = 231.2$, df = 180, $P = 0.01$). In this analysis, all 20 items had ordered thresholds (Fig. 5) and nonsignificant chi-square P values after Bonferroni adjustment. Item fit was within ±2.5 for 10 items. Reliability was high with PSI and Cronbach alpha values ≥0.97. Regarding local dependency, six pairs of items had residual correlations more than 0.30. After subtests were performed, PSI values did not change, and Cronbach alpha values were 0.96. DIF was evident for two items for age-group and two items for gender. When items with DIF were split by the relevant patient characteristic, correlations between the original and split person locations did not impact scoring (r = 1.00). The person-item threshold distribution shows the scale had a good range of measurement; 90.1% of participants scored on the scale (Fig. 4). Floor (1.5%) and ceiling (8.4%) effects were low.

Construct Validity

As hypothesized, scores for the two item libraries, the original 10-item scales,[6,9] and the three short-form appearance scales were incrementally lower ($P \leq 0.001$) as the

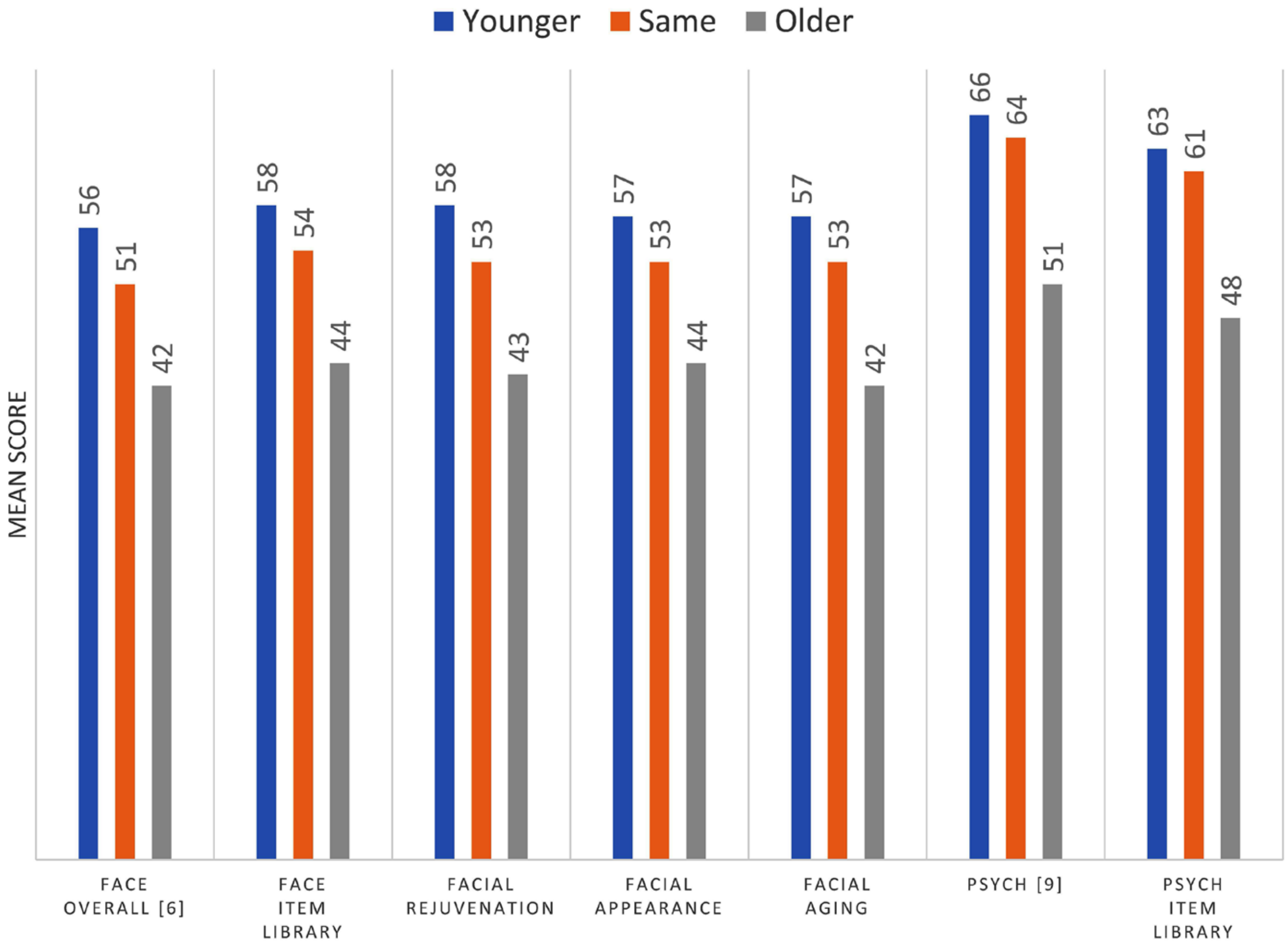

Fig. 6. Mean FACE-Q Aesthetics scores based on how much younger or older someone looks compared with their actual age.

degree to which participants were more bothered by lax facial skin, looked older than their actual age, and had more of their treatment wear off (Figs. 6–8). Scores were also incrementally lower as Merz Assessment Scale scores[40] for the severity of dynamic and static lines increased. [**See table, Supplemental Digital Content 7**, which displays FACE-Q Aesthetics scores (Mean, SD) for self-reported depth of facial lines based on Merz Assessment scales. **http://links.lww.com/PRSGO/D151**.]

Test–Retest Reliability

Of the 118 participants who completed the test–retest reliability, three participants who completed the test–retest reliability after 14 days were excluded. In addition, eight and six participants who reported a change in satisfaction with face and psychological function, respectively, were excluded.

The number of extreme cases was 10 in the face item library, eight in the psychological item library, 10 in the Face Overall scale,[6] and seven in the Psychological Function scale.[9] After these exclusions, ICC values without extremes were more than 0.80, and with extremes were more than 0.70 (Table 4).

DISCUSSION

Our qualitative study elicited new appearance and psychological concepts in the context of minimally invasive aesthetic treatments. These new concepts effectively extended the range of measurement of the FACE-Q Aesthetics Face Overall[6] and Psychological Function[9] scales. Our study adds to previously published findings showing strong psychometric performance of both original FACE-Q Aesthetics scales.[6,9] The qualitative phase of our study supported content validity for 52 face and 22 psychological items. In the quantitative phase, a range of evidence was used to identify the best subset of items to retain that also included the original Face Overall[6] and Psychological Function[9] items. Taken together, the psychometric evidence supports the reliability and validity of the final 42 face and 20 psychological items chosen. Data from 1369 people evidenced good fit to the Rasch model with little evidence of DIF by age, gender, or country. The psychometric tests support the overall quality of the item libraries, and the quality of three appearance short-form scales created as examples of how the item libraries could be used. Validation of a PROM is an ongoing process that involves the accumulation of evidence over time.[38] Future

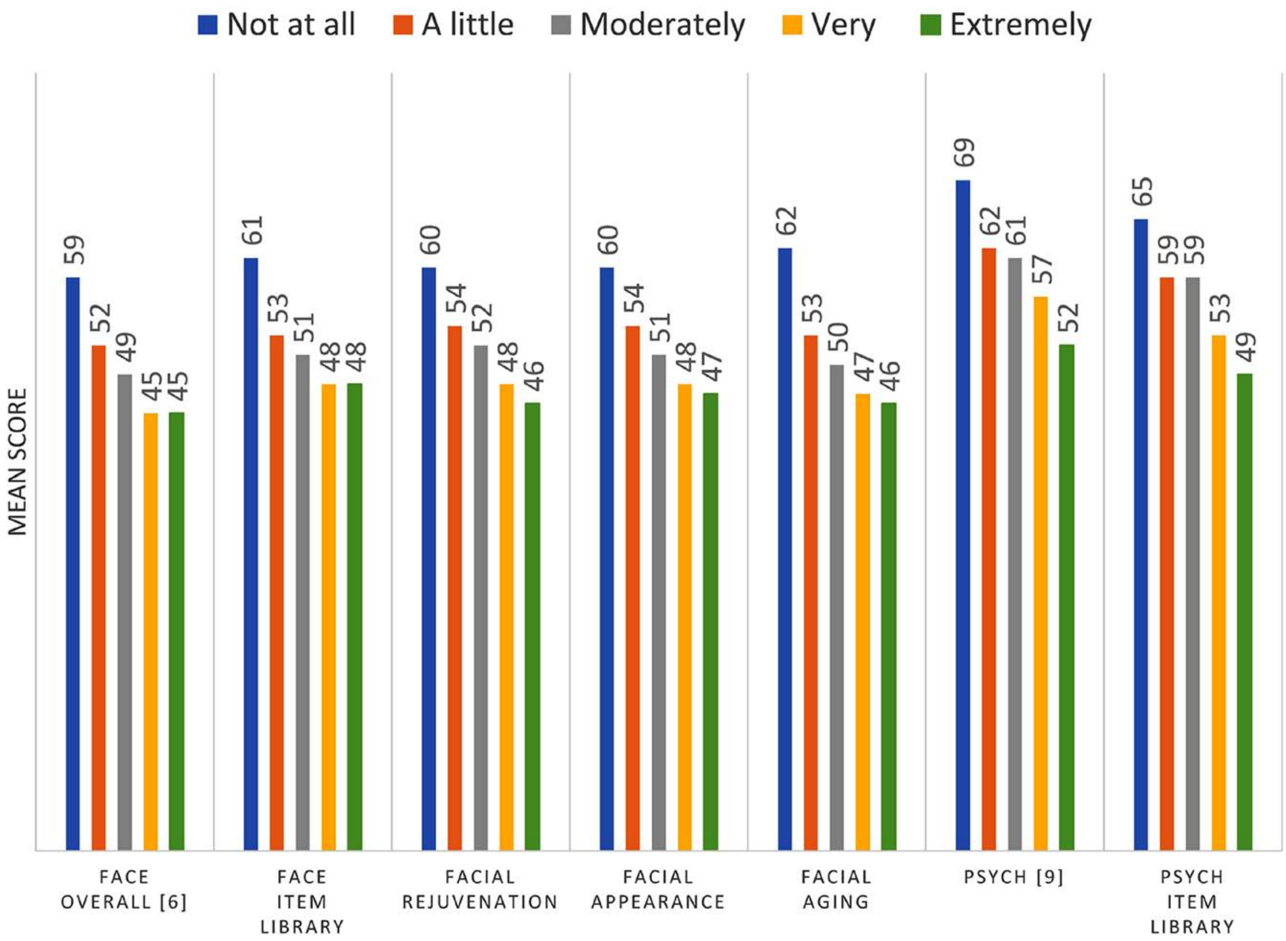

Fig. 7. Mean FACE-Q Aesthetic scores based on how much participants were bothered with facial skin laxity.

directions for this research include exploring responsiveness and establishing minimally important differences.

Our study's use of a modern psychometric approach has important advantages over the traditional psychometric approach.[32] Scales designed using RMT analysis can be used as an item library or item bank depending on the intended use. More specifically, scores for short-form scales can be calibrated in relation to the complete set of items using the Rasch model (ie, item bank approach), or stand-alone scoring can be created (ie, item library approach). For the item bank approach, scores from the original scales[6,9] can be directly equated via a crosswalk that links their respective scoring algorithms. In addition, scales developed using Rasch analysis are amenable to computer adaptive tests (CATs). A CAT uses an algorithm to shorten the number of items a person needs to complete in a PROM by selecting the next most relevant items based on the answers provided. CATs are appealing in clinical care as they can substantially reduce respondent burden while maintaining a high level of accuracy.[41]

This study has some limitations. First, we aimed to include a broad range of treatments but recognize that some treatments were represented by a smaller number of participants than others. Second, the qualitative sample did not include participants from the United Kingdom. However, we confirmed content validity for the United Kingdom by having 40 residents complete the cognitive survey before the field test. Furthermore, we did not find evidence of DIF by country for any item. Additional research to validate the FACE-Q item libraries in other countries and languages is warranted. Third, participants in our study may not reflect the general population of the United States, Canada, or the United Kingdom, or reflect people who undergo facial aesthetic treatments.[17] Fourth, our sample focused on minimally invasive treatments. Future research could establish the applicability of the item libraries for surgical treatments. Fifth, the use of online platforms for research involves people who self-select to take part and are paid for their involvement. We used the Prolific platform as it has been shown to be high quality compared with other online platforms.[42,43] Finally, all data collected in our study were self reported, including the Merz Assessment Scale scores,[40] which were not verified clinically.

To conclude, the new item libraries provide an innovative approach to measuring appearance and psychological well-being in the context of minimally invasive treatments. This PROM approach allows end-users to customize fit-for-purpose short-form scales for use in clinical trials and

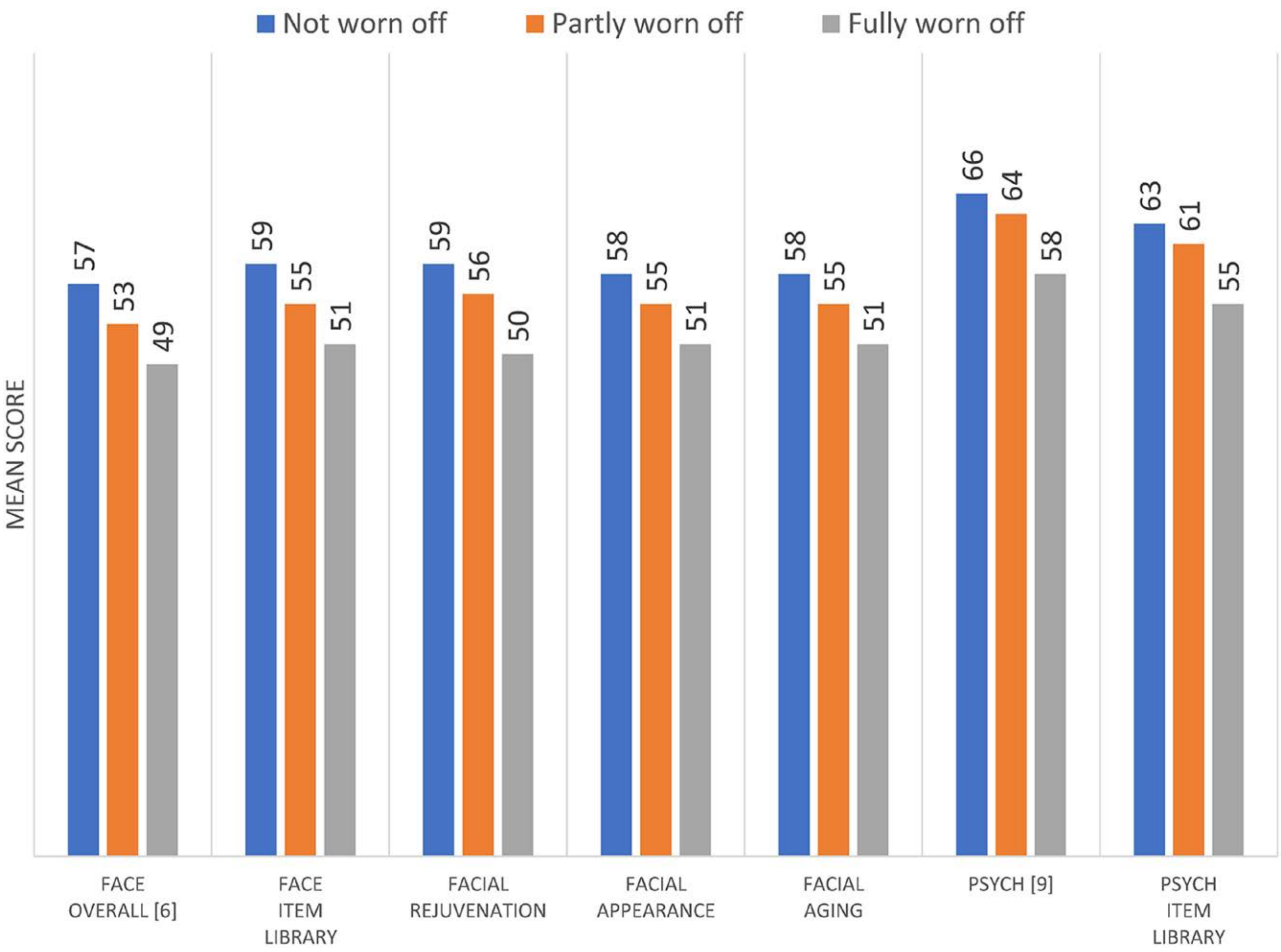

Fig. 8. Mean FACE-Q Aesthetics scores based on self-report of the amount the facial aesthetic treatment has worn off.

clinical practice. A license to use FACE-Q Aesthetics is available through https://qportfolio.org/face-q/aesthetics/.

Anne Klassen, DPhil
3N27, Health Sciences Center
1280 Main Street W
Hamilton, Ontario
Canada L8S 4L8
E-mail: aklass@mcmaster.ca

DISCLOSURES

Drs Klassen, Cano, and Pusic are co-developers of the FACE-Q Aesthetics patient-reported outcome measure and as such receive a share of license revenues as royalties based on their institutions' inventor sharing policy. Stefan Cano is the CSO of Modus Outcomes, a Division of Thread. Anne Klassen provides research consulting services to the pharmaceutical industry through EVENTUM Research. All the other authors have no financial interest to declare in relation to the content of this article.

REFERENCES

1. US Food and Drug Administration. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. Available at https://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf. Accessed September 27, 2023.
2. Calvert M, Kyte D, Price G, et al. Maximizing the impact of patient reported outcome assessment for patients and society. *BMJ.* 2019;364:k5267.
3. Black N. Patient reported outcome measures could help transform healthcare. *BMJ.* 2013;346:f167.
4. Nelson EC, Eftimovska E, Lind C, et al. Patient reported outcome measures in practice. *BMJ.* 2015;350:g7818.
5. Klassen AF, Cano SJ, Scott A, et al. Measuring patient-reported outcomes in facial aesthetic patients: development of the FACE-Q. *Facial Plast Surg.* 2010;26:303–309.
6. Pusic A, Klassen AF, Scott AM, et al. Development and psychometric evaluation of the FACE-Q Satisfaction with Appearance Scale: a new patient-reported outcome instrument for facial aesthetics patients. *Clin Plast Surg.* 2013;40:249–260.
7. Panchapakesan V, Klassen AF, Cano SJ, et al. Development and psychometric evaluation of the FACE-Q Aging Appraisal Scale and patient-perceived Age Visual Analog scale. *Aesthet Surg J.* 2013;33:1099–1109.
8. Klassen AF, Cano SJ, Scott AM, et al. Measuring outcomes that matter to face-lift patients: development and validation of FACE-Q appearance appraisal scales and adverse effects checklist for the lower face and neck. *Plast Reconstr Surg.* 2014;133:21–30.
9. Klassen AF, Cano SJ, Schwitzer J, et al. FACE-Q scales for health-related quality of life, early life impact and satisfaction with outcomes and decision to have treatment: development and validation. *Plast Reconstr Surg.* 2015;135:375–386.

10. Klassen AF, Cano SJ, East CA, et al. Development and psychometric evaluation of the FACE-Q scales for patients undergoing rhinoplasty. *JAMA Facial Plast Surg.* 2016;18:27–35.
11. Klassen AF, Cano SJ, Schwitzer JA, et al. Development and psychometric validation of the FACE-Q skin, lips, and facial rhytides appearance scales and adverse effects checklists for cosmetic procedures. *JAMA Dermatol.* 2016;152:443–451.
12. Klassen AF, Cano SJ, Grotting JC, et al. FACE-Q Eye Module for measuring patient-reported outcomes following cosmetic eye treatments. *JAMA Facial Plast Surg.* 2017;19:7–14.
13. Ottenhof MJ, Veldhuizen IJ, Hensbergen LJV, et al. The use of the FACE-Q Aesthetic: a narrative review. *Aesthetic Plast Surg.* 2022;46:2769–2780.
14. Hoffman L, Fabi S. Look better, feel better, live better? The impact of minimally invasive aesthetic procedures on satisfaction with appearance and psychosocial wellbeing. *J Clin Aesthet Dermatol.* 2022;15:47–58.
15. Gallo L, Kim P, Yuan M, et al. Best practices for FACE-Q Aesthetics research: a systematic review of study methodology. *Aesthet Surg J.* 2023;43:NP674–NP686.
16. Food and Drug Administration. MDDT summary of evidence and basis of qualification decision for FACE-Q | Aesthetics. Available at https://www.fda.gov/media/157956/download. Accessed September 27, 2023.
17. American Society of Plastic Surgery. 2009 plastic surgery statistics. Available at https://www.plasticsurgery.org/news/plastic-surgery-statistics?sub=2009+Plastic+Surgery+Statistics. Accessed September 27, 2023.
18. Choppin B. Item banking using sample-free calibration. *Nature.* 1968;219:870–872.
19. Massof RW, Ahmadian L, Grover LL, et al. The Activity Inventory: an adaptive visual function questionnaire. *Optom Vis Sci.* 2007;84:763–774.
20. Regnault A, Pompilus F, Ciesluk A, et al. Measuring patient-reported physical functioning and fatigue in myelodysplastic syndromes using a modular approach based on EORTC QLQ-C30. *J Patient Rep Outcomes.* 2021;5:60.
21. Piccinin C, Basch E, Bhatnagar V, et al. Recommendations on the use of item libraries for patient-reported outcome measurement in oncology trials: findings from an international, multidisciplinary working group. *Lancet Oncol.* 2023;24:e86–e95.
22. Rose M, Bjorner JB, Gandek B, et al. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *J Clin Epidemiol.* 2014;67:P516–P526.
23. Regnault A, Willgoss T, Barbic S. On behalf of the International Society for Quality of Life Research Mixed Methods Special Interest Group. Towards the use of mixed methods inquiry as best practice in health outcomes research. *J Patient Rep Outcomes.* 2018;2:19.
24. Aaronson N, Alonso J, Burnam A, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res.* 2002;11:193–205.
25. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 1—eliciting concepts for a new PRO instrument. *Value Health.* 2011;14:967–977.
26. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. *Value Health.* 2011;14:978–988.
27. Thorne S, Kirkham SR, MacDonald-Emes J. Interpretive description: a noncategorical qualitative alternative for developing nursing knowledge. *Res Nurs Health.* 1997;20:169–177.
28. Pope C, Ziebland S, Mays N. Qualitative research in health care. Analysing qualitative data. *BMJ.* 2000;320:114–116.
29. Sandelowski M. Theoretical saturation. In: Given LM, ed. *The Sage Encyclopedia of Qualitative Methods.* Thousand Oaks, Calif.: Sage; 2008;1:875–876.
30. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42:377–381.
31. Andrich D, Sheridan BS, Luo G. *RUMM2030Plus: Rasch Unidimensional Models for Measurement.* Perth, Western Australia: RUMM Laboratory. Available at www.rummlab.com.au. Published 2021.
32. Andrich D, Marais A. *Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences.* Singapore: Springer Texts in Education, Springer; 2019.
33. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess.* 2009;13:1–177.
34. Christensen KB, Makransky G, Horton M. Critical values for Yen's Q3: identification of local dependence in the Rasch model using residual correlations. *Appl Psychol Meas.* 2017;41:178–194.
35. Cleanthous S, Bongardt S, Marquis P, et al. Psychometric analysis from EMBODY1 and 2 clinical trials to help select suitable fatigue PRO scales for future systemic lupus erythematosus studies. *Rheumatol Therapy.* 2021;8:1287–1301.
36. Andrich D, Hagquist C. Real and artificial differential item functioning. *J Edu Behav Statist.* 2012;37:387–416.
37. Andrich D. An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern. *Educ Res Perspect.* 1982;9:95–104.
38. Nunnally JC. *Psychometric Theory.* 3rd ed. New York, N.Y.: McGraw-Hill; 1994.
39. Prinsen CA, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27:1147–1157.
40. Stella E, Di Petrillo A. Standard evaluation of the patient: the Merz Scale. In: Goisis M, eds. *Injections in Aesthetic Medicine.* Milan, Italy: Springer; 2014.
41. Harrison CJ, Apon I, Ardouin K, et al. The development, deployment, and evaluation of the CLEFT-Q Computerized Adaptive Test: a multimethods approach contributing to personalized, person-centered health assessments in plastic surgery. *J Med Internet Res.* 2023;25:e41870.
42. Peer E, Rothschild D, Gordon A, et al. Data quality of platforms and panels for online behavioral research. *Behav Res Methods.* 2022;54:1643–1662.
43. Douglas BD, Ewell PJ, Brauer M. Data quality in online human-subjects research: comparisons between MTurk, Prolific, CloudResearch, qualtrics, and SONA. *PLoS One.* 2023;18:e0279720.